HIV Alignments, Database Searches, and Structure Predictions

Gerald Myers and Andrew Farmer

MS K710, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Analyses of more distantly related HIV and SIV sequences that take as their point of departure an alignment, either of the nucleic acid or amino acid sequences, will only be as sound as the alignment, which is itself an hypothesis. For this reason, many sequence analyses are conducted over only "unambiguously alignable" stretches of sequence, typically stretches for which the similarities are 50% or greater, and the information in the more varied regions (similarities less than 30%) is lost to the analysis. This is unfortunate insofar as structural information is typically derived from distantly related, not closely related, sequences. Given the diversity of primate immunodeficiency viral (PIV) sequences, alignments in the database publications up to this year were constrained of necessity to highly related subgroups of viral types, which were then "apposed" by eye. These safe, but restricted, alignments could not support some analyses based upon the entire set of PIV sequences. This year, in Parts I, II and III, we have brought into play a new alignment strategy that holds some promise for simultaneously and objectively aligning all members of the primate immunodeficiency virus family.

An additional reason for exploring new approaches to HIV and SIV sequence alignment concerns computational time: multiple alignment, viewed as an extension of pairwise alignment, requires time proportional to the average length of the sequences raised to power *k*, where *k* is the number of sequences being aligned. In the approach described below, sequences are aligned in time linearly proportional to *k*, the number of sequences. There are further advantages to the method. Most alignment strategies are "progressive", which is to say that the alignment unfolds from the pairs of most similar sequences to the pairs of most dissimilar sequences; the essence of this approach is captured by Doolittle's dictum—"once a gap always a gap" [1]. McClure and coworkers critique twelve different alignment methods, most of which are progressive, according to their abilities to correctly identify ordered series of motifs in highly divergent proteins that have been experimentally studied [2]. They find that no single approach is superior to all others, and most are time-consuming. Some of the newer multiple alignment programs are intentionally not progressive, partly for the reason that progressive alignments may be trapped by local optima, partly because phylogenetic inferences are implicitly assumed. The Hidden Markov Method (HMM) approach, which we utilize and describe herein, is not progressive; instead, it emphasizes position-specific probability distributions of character states, hence a gap in one portion of the alignment may be scored differently than a gap in another portion of the aligment. Most alignment programs have position-independent scoring schemes, which are unrealistic in the case of most proteins since they are composed of both conserved regions and indel-rich variable regions. As HMM is centered upon the columns of information, this approach, referred to as a "generalized profile", is indifferent to the relatedness of pairs of sequences [3,4]. On the other hand, because a probability distribution over the 20 amino acids must be constructed at each position in the sequence, HMM is employed with large sets (40 or more taxa) of highly divergent sequences such as is seen with HIVs and SIVs. Furthermore, an assumption of independence between sites is made. The latter assumption is not universally defensible, however neither is it new to sequence analyses.

At this point in the discussion, it will be helpful to briefly recall what a Markov chain is and in what sense the sequence alignment problem is said to be "hidden." A first order Markov process is one in which the state at time t is a probabilistic function of the state at time $t - 1$. We *think* that this is a reasonable assumption for viral sequence evolution. Many phenomena describable by a Markov process are observable, *e.g.*, changes in weather where a complete history of weather is on record [5]; when the changes of state are themselves not observable, the process is said to be hidden. In the case of HIV sequences, we do not possess a history of the intermediate states through which sequences have evolved. Fitch's notion of a *covarion*, the set of concomitantly variable codons, can illustrate the hiddeness: we imagine that evolutionary changes shift the makeup of the covarion, at one time a position be invariant, at another time being variable; one state representing one covarion, another state representing another covarion. These states are not observed. The overall problem is to formally arrive at a probabilitistic model that will satisfactorily account for the sequences that are known. Specific aspects of the problem, in addition to facilitating alignments, are to derive a suitable "consensus" sequence, to classify a sequence (is it distantly related or totally unrelated?), to support structural analysis, and to gain insight into the "hidden" evolutionary process.

Hence the HMM approach leads to a model for the sequence set that has been analyzed. An example of an HMM-generated model, or so-called architecture, is shown on the cover of this year's compendium. With subsequent database searches, this model—in the form of a "most likely" sequence, or *discriminator*—embodies all of the information contained in the data set, not merely one particular sequence. As we shall see, this consensus-like sequence is useful for database searching (below and accompanying Part IV section concerned with Molecular Mimicry). We have also coupled the HMM-generated model to prediction analyses of protein structure using an array of contemporary algorithms. Eventually, the sequence alignment and the structure prediction will become intertwined in an effort to optimize the alignment.

Be forewarned that the method is still in its infancy and that our utilization is at the most elementary level: refinement of the approach, especially to extend the analysis to nonprimate lentiviruses and other retroviruses, entails extensive parameterization studies such as those being undertaken by McClure and coworkers [6]. Issues such as optimum model length, size of training set, etc. are not taken up in this analysis, which was restricted to just primate lentiviral sequences.

In the following text, we first describe in some detail the HMM approach to alignment as we have applied it in this compendium, especially in Parts I and II, then we turn to discussions of database searching and protein structure prediction.

A. MULTIPLE SEQUENCE ALIGNMENT OF HIVS and SIVS USING HMM

The Hidden Markov Model (HMM), as it has been applied to sequence analysis, has many similarities to what is called a "profile" [7,8] in terms of the information that it captures concerning a set of related sequences. In a sense, each can be thought of as an extended consensus sequence in which the information retained at each position includes the frequency with which each possible base or amino acid residue is seen in the sequence set at that position. The HMM is constructed from a number of successive nodes generally corresponding to the columns of positional homology of an alignment; each of these nodes contains a match state, an insert state and a delete state (see figure on the cover of this compendium). Match states correspond to simple amino acid (or base) substitutions; insert and delete states are self-explanatory. Associated with each of the states in the model are vectors of probabilities that specify the likelihood with which the system might pass to each member of the set of next possible states; these are referred to as transition probabilities. Also associated with match states and insert states are vectors of probability specifying the likelihood that the system will generate or "emit" each possible amino acid or nucleotide when in that state; delete states allow for the possibility that a sequence not have a character in a certain column. Altogether, there will be three probability matrices to describe the model—the transition matrix, the emission matrix, and the initial state matrix.

The resultant architecture of the HMM allows one to establish a correspondence between the characters of a given sequence and the states of the model. The succession of the characters in the sequence will thus determine a path through the states of the model, and associated with this path will be a likelihood determined both by the probabilities of transition between successive states and the probability that each state has for generating the character that has been assigned to it. Provided that all the probabilities in the model, including both transition and emission probabilities, are non-zero, then each path through the model that is permissible according to the rules governing transitions from one node of the model to the next will have a non-zero probability of generating the given sequence. The task of finding the optimal path through the model for a given sequence, i.e. the path with the highest likelihood, can be thought of as aligning the sequence to the model, and may be solved using dynamic programming techniques.

The most important differences between the profile and the HMM lie not in the resultant information structures, but in the means by which these structures are generated from the sequence data. As with an ordinary consensus sequence, the profile is generated from a set of sequences whose alignment has been determined by some independent means. The parameters for describing an HMM can also be derived from a given alignment in this manner. More importantly, however, there exists an algorithm for HMMs that allows one to determine the parameters of the model having the highest likelihood (at least within the neighborhood of the initial model) given a set of unaligned sequences. This approach is quite similar to certain techniques used in connection with artificial intelligence applications, and is known as “training” the model. The general procedure for training makes use of an Expectation-Maximization algorithm, of which there are several [3,4,6].

Speaking generally, the algorithm used in training the parameters of an HMM involves an iterative approach that uses an initial model to estimate an alignment of the given set of sequences, then uses this alignment to re-estimate the model, and so on until the estimates converge to an optimum. For example, if we are given a set of protein sequences that are thought to be related, a good estimate for an initial model can be made by using the frequency distribution of amino acids in the unaligned set as a vector of probabilities assigned to all the match states and insert states of the model; transition probabilities between the states of the initial model can be assigned arbitrarily, or using a prior assessment of the relative frequencies of indel events. All of the sequences in the given set will now be aligned in turn to the model, finding the path through the model that maximizes the likelihood for the given sequence; by aligning all the sequences in the set to the model in this pairwise fashion, one transitively defines a multiple sequence alignment of the sequences to one another. The multiple sequence alignment thus created can be used for an estimate of the parameters of the HMM, by counting the frequency of occurrence of each amino acid at each position of the alignment and the frequency of indel events across the alignment. This adjusted HMM then serves as a model for another round of alignment, and so on. It can be shown that this process is guaranteed to converge to a local maximum of the likelihood function.

To address the problem of guaranteeing convergence to a global maximum for this function, a variation of the simulated annealing algorithm can be applied at each step of the iterative algorithm; this basically allows a stochastically generated sub-optimal alignment to be chosen for the re-estimation of the model’s parameters, where the sub-optimality of the alignment decreases to zero with successive iterations of the re-estimation procedure.

As should be clear from the preceding discussion, the model can be used to generate a multiple sequence alignment of sequences, including sequences not belonging to the set used to train the parameters of the model. A distinct advantage to using the HMM over the standard dynamic programming algorithm for multiple sequence alignments is that since one is really performing a set of pairwise comparisons of the sequences to the model, the time and memory requirements increase only linearly with the number of sequences, as opposed to exponentially with dynamic programming. It follows that it is relatively easy to add a new sequence to the alignment and rebuild the model; experimentally-derived information can also be added to the model (known as *priors*) with relative ease.

A good general introduction to the basic ideas of HMMs (not oriented, however, toward sequence analysis) is reference 5 below. We have employed the HMMER implementation that is publicly available [9–10](*eddy@genetics.wustl.edu*). Another HMM suite that can be obtained is SAM (*http://www.cse.ucsc.edu/research/compbio/sam.html*). The SAM Web site contains a number of links to papers concerned with HMMs and sequence analysis. These programs were originally applied to highly studied data sets, “validation” sets [3,4,6] (globins, EF-hand proteins, kinases, and proteases), for which extensive experimentally-based data were available to help assess the alignment results. With HIV and SIV sequences, the results of the approach must be critiqued by scutinizing motifs—do cysteines, cleavage sites, and potential glycosylation sites in envelope align, for example? This can be problematic as it is not preordained, for instance, that all sequons need align [11]. Another approach to critiquing the HMM-generated alignment involves construction of blocks using representative PIV sequences (described below). Finally, it is necessary to compare scores from matching individual HIV and SIV sequences to the “most likely”, or *discriminator*, sequence, the HMM equivalent of a consensus: as we shall report in the section to follow, our model has been “overfitted” to HIV-1 sequences; nevertheless, the boundaries and motifs appear to be satisfactorily aligned.

Approximately 400 HIV and SIV sequences were trained using HMMER version 1.8. The frequency distribution of amino acids in the unaligned HIV and SIV sequences was used as input in place of the default distribution derived from the PIR database. We should first consider the success or failure of the approach with respect to identifying motifs. The envelope protein alignment of an HIV2 sequence (SBL/ISY) to the HMM consensus is instructive in this regard (figure 1): 1) cysteines, denoted by ’*’, and potential N-linked glycosylation sites, denoted by ˆˆˆ, are aligned as we might expect them to be aligned when doing ordinary alignment assisted by manual input; 2) four noncontiguous residues found in certain HIV-1 subtype B sequences to be essential for CD4 interaction, D, E, W and D (see pp. II-1,2 of NOV 95), are aligned in almost all HIVs and SIVs (tryptophans are in general highly conserved); 3) the gp120/gp41 cleavage site is aligned, as we expect, although an alternative cleavage site may be used in some HIV-2s.

Further confirmation of the alignment comes from BLOCKS analysis: using representative sequences from the Part II HMMER-generated alignments, blocks—gapless arrays of multiply aligned conserved sequences—were constructed using the BLOCKMAKER program [12,13] (http://www.blocks.fhcrc.org). Two different programs are employed by BLOCKMAKER, the MOTIF program and the GIBBS Sampler program. Boundaries for the created blocks were highly conserved in the HMM alignments (Part II), however blocks based on envelope sequences largely coincide with conserved domains, hence other motifs (cysteines, glycosylation sites) must provide the main support for alignment over the more varied regions.

A difficulty encountered by the HMM (and every) alignment method is large indels; we have the least confidence in those. To the extent that these stretches may have arisen through acquisition of genetic material, they may not be intrinsically alignable as they may not be homologous. The reader may want to compare the V1-V2 region alignment in figure 1 with one manually created by Lamers et al. [14]. Alignments in previous publications may be “safer”, but they are also more constrained and less informative because they were executed over just highly related sequences. Since the different alignments have different applications, both are made available on the Web site (*http://hiv-web.lanl.gov*).

Nucleotide sequence alignments, by this approach, were produced from the HMM-generated amino acid sequence alignments. The nucleotide alignment was then subjected to HMMER and a nucleotide-specific model was obtained. This approach follows in a general way the approach taken in earlier database publications that were based on the PIMA algorithm of Smith and Smith.

B. DATABASE SEARCHING USING AN HMM APPROACH

An important application of the HMM-generated model is in the discrimination of related sequences from non-related sequences. This is especially useful in connection with database searching. Associated with each sequence in the database is a probability, a log-odds score analogous to a BLAST score, with which the sequence could be generated by the given model. With the HMMER algorithm [10],

$$\text{score} = \log_2 \frac{P(S_i|M)}{P(S_i|R)}$$

where the alignment of each sequence in the database, S_i, is compared to both the HMM generated model, M, and a random model, R. The latter should have the same amino acid composition as the database at large and it should be as likely, *a priori*, as M. The log-odds score corrects for sequence length. A score greater than zero has a better than even chance of being significant, however, as a rule of thumb, a score must be about 20 or more to be deemed significant [9]. The distribution of likelihood scores for all the sequences in the database will provide a measure of discrimination between similar and non-similar sequences. Using the HMM for database searching has the advantage of utilising a great deal more of the information available for a family of sequences than can be captured by query techniques that force one to use only one sequence from the family or, at best, a standard consensus sequence as a query against the database.

Figure 1. Comparison of an HIV-2 Env Sequence to the HMM Model

```
                                             *
HMMER Model       *MRVKGIQRNWQHWWRWGTMLLGMLMICSAAENLWVTVYYGVPVWKEATT
                   M     ++ +  ++ + ++++++L++C      +VTV+YGVPVWK+A++
HIV-2SBL/ISY    1  M----SGKIQ--LLVAFLLTSACLIYC----TKYVTVFYGVPVWKNASI    39
                     *                   *             ^^^
                  TLFCASDAKAYDTEVHNVWATHACVPTDPNPQEIVLENVTENFNMWKNNM
                  +LFCA +++      +++W+T++C+P+++++QEI+L NVTE+F++W+N +
               40 PLFCA-TKN------RDTWGTIQCLPDNDDYQEIPL-NVTEAFDAWDNIV    81
                                    *      *    *   ^^^  ^^^^^^
                  VEQMHEDIISLWDQSLKPCVKLTPLCVTLNCTDWNaT..NTTNTTn....
                  +EQ+ ED+++L+++S+KPCVKLTPLCVT+NC+  +     TT+ ++
               82 TEQAVEDVWNLFETSIKPCVKLTPLCVTMNCNASTESAVATTSPSGPDMI   131
                                       ^*^ ^^^
                  .............GMEKGEMKNCSFNMTTEIRDKKQKEYALFYKLDVVPI
                               G+ ++ M+ C+FNMT++  DKK+++ +++Y+ D V++
              132 NDTDPCIQLNNCSGLREEDMVECQFNMTGLELDKKKQYSETWYSKD-VVC   180
                    ^^^^          *^^^     *            *         *
                  DNNNTS.....YRLINCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCN
                  + +N++     Y+ ++CNTSVIT++C+K+++++++++YCAP+GF +L+CN
              181 ESDNSTDRKRCYM-NHCNTSVITESCDKHYWDAMRFRYCAPPGFVLLRCN   229
                      ^^^   * ^^^   *              ^^^           ^^^
                  DKKFNGTGP.CKNVSTVQCTHGIKPVVSTQLLLNGSLAEEEIVIRSENFT
                  D++++G++P C++V++++CT++++++ ST+L++NG++AE++++I+++++
              230 DTNYSGFEPNCSKVVASTCTRMMETQPSTWLGFNGTRAENRTYIYWHGR-   278
                                   ^*^   ^^^
                  NNAKTIIVQLNE.SVEINCTRPNNNTRKSIHIGPGQAFYTTGDIIGDIRQ
                  +N +TII+++++ +++I C RP N+T+++I++++G+ F+++  I+  +RQ
              279 DN-RTIISLNKYYNLTILCRRPENKTVVPITLMSGRRFHSQKIINKKPRQ   327
                                                              CD4 CD4
                    *^^^    ^^^                  ^^^     ^^^    | |
                  AHCNISRTKWNNTLQQIVAQTLKKL.REHFGNKT..IIFNQSS.GGDPEI
                  A+C ++++ W +++Q+ V+QTL+K+ R++++N+T  I+F+++   +DPE+
              328 AWCRFKGE-WREAMQE-VKQTLVKHPRYKGTNDTNKINFTAPEKDSDPEV   375
                         *      *^^^   ^^^   ^^^  ^^^  ^^^    *
                  T.MHSFNCGGEFFYCNTTWLFNSTWn.NgTWSNNTEGNDTITLPCRIKQI
                    M+ +NC+GEF+YCN+TW F+ +W+ N+T+       +++++PC+I QI
              376 AYMW-TNCRGEFLYCNMTW-FL-NWVENKTG------QQHNYVPCHIEQI   416
                   CD4                           CD4
                     |                 *  ^^^      |   ^^^ ^^^
                  INMWQEVGKAMYAPPIEGQIRCSSNITGLLLTRDGGNNNT.NETF.RPGG
                  IN+W++VGK++Y+PP+EG+++C+S++T++++++D    N  N+TF
              417 INTWHKVGKNVYLPPREGELSCESTVTSIIANIDVDGDNRTNITFS----   462
                                                         gp120/gp41 in HIV-1
                  G.DMRDNWRSEL..YKYKVVKIEPLGVAPTKAKRRV..VQREKRAV.GIG
                    ++++++R+EL  YK+  V+ +P+G+APT  KR++  ++R+KR+V ++G
              463 -AEVAELYRLELGDYKL--VEVTPIGFAPTAEKRYSSAPGRHKRGVLVLG   509
                  AMFLGFLGAAGSTMGAASMTLTVQARQLLSGIVQQQNNLLRAIEAQQHML
                    FLGFL +AG +MGA+S+TL++Q+R+L  GIVQQQ++LL+++++QQ+ML
              510 --FLGFLTTAGAAMGARSLTLSAQSRTLFRGIVQQQQQLLDVVKRQQEML   557
                                                 *     *       ^^^
                  QLTVWGIKQLQARVLAVERYLKDQQLLGIWGCSGKQICHTTVPW.NSSW.
                  +LTVWG+K+LQARV+A+E+YL DQ+ L++WGC+++Q+CHTTVPW N+++
              558 RLTVWGTKNLQARVTAIEKYLADQARLNSWGCAFRQVCHTTVPWVNDTLT   607
                   ^^^      ^^^         ^^^
                  SNKSLDPIWNNMTWMEWEREIDNYTaNIYtLIEESQNQQEKNEQELLELD
                        P WNNMTW+EWE +I++++ANI++++E++Q+QQEKN++EL++L+
              608 ------PEWNNMTWQEWEHKIRFLEANISESLEQAQIQQEKNMYELQKLN   651
                  KWASLWNWFDITNWLWYIKIFIMIVGGLIGLRIVFYVLSIVNRVRQGYSP
                  +W++++NWFD+T+W++YI++++MIV+G+++LRIV+YV+++++R+R+GY+P
              652 SWDVFGNWFDLTSWIKYIQYGVMIVVGIVALRIVIYVVQMLSRLRKGYRP   701
                  LSSSPP.Y.FQTHIPHPRGPDRPEGIEEEGGEQDRDRSWRW...VNGFLA
                  ++SSPP Y +Q+HI+++++++ +E++EE++G+  + RSW+W   +++F
              702 VFSSPPGYIQQIHIHKDWEQPDREETEEDVGNDVGSRSWPWPIEYIHF--   749
                           *
                  LIWDDLRSLC.LFSYHRLRDLILIVA.RIVELLGRRGWEALK.YWWNLLQ
                  LI+ ++R+L+ L++++R+++++L+ + ++      R+W++LK    ++LQ
              750 LIRLLIRLLTRLYNSCRDLLSRLYLILQPL-----RDWLRLKA---AYLQ   791
                  Y...WSQELKNSAVSLLNATAIAVAEWTDRWIEVQRICRAILHIPRRIRQ
                  Y   W+QE++++ ++ +++T +++A+ + +W++++RI+R+IL++PRRIRQ
              792 YGCEWIQEAFQALARVTRET-LTSAG-RSLWGALGRIGRGILAVPRRIRQ   839
                  GLERALL*
                  G+E+ALL
              840 GAEIALL                                              846
```

The HMMER program, used herein, employs the Smith-Waterman (S-W) algorithm for optimizing local alignments. As with BLAST, both identities and equivalencies (conserved amino acid replacements) are scored; unlike BLAST, which is a heuristic search program, the S-W algorithm reports only the best match between the HMM consensus and a given database entry. A general discussion of these somewhat different approaches to database searching is found in reference [15]. Compositional bias may be present in the outcome of a database search, which is to say that spurious matches may arise by virtue of similar compositions only.

The comparison shown in figure 1 of a particular HIV-2 sequence to the HMM model consensus was taken from an S-W/HMMER search of the protein database using the HMM model for envelope. The score (corrected for the model length and the length of the target sequence) was highly significant, 1676, however HIV-1 scores uniformly hovered between 2700 and 2800, which tells us that the model was overfitted to HIV-1s. Scores among HIV-1s of different genetic subtypes did not significantly differ, which was satisfying, but HIV-2 and SIV scores were uniformly lower. The *maximum discrimination* option of HMMER, in contrast to the default *maximum likelihood* improves this situation somewhat [9–10]. Note in figure 1 that the reduction in score for the HIV-2 is mostly due to amino acid differences regarded to be conservative (indicated with a'+'). Although the probability distributions at the various homologous sites did not have adequate representation from the HIV-2 sequences, the alignment, as such, is reasonable. The scores over just the envelope gp41 were closer; and the scores for gag protein were much closer, 1759 for an HIV-1, 1478 for an HIV-2, showing that the fit is more inclusive.

Parallel S-W/HMMER envelope searches were conducted using a database in which HIV and SIV sequences were filtered out. The only non-zero matches to nonprimate lentiviral sequences involved Visna and its close relatives, OMVVSA and CAEV. The log-odds scores for these were less than 10; for example, from the Env gp41,

```
HMM Consensus    NNMTWMEWEREIDNYTaNIYtLIEES
                 +N TW++WERE   Y +N + L+ ES
     CAEV Env    DNCTWQQWERELQGYDGNLTMLLRES
```

The score for this match was 5.55, suggesting that the cutoff value of 20 should not be too rigidly applied. The highest match score in this search, 12.15, belongs to a horse skeletal muscle sodium channel alpha-subunit, which illustrates the possibility of compositional bias:

```
HMM Consensus    NTTWLFNSTWn.NgTW.SNNTEG.ND
                 NTTW  N TW+ N+TW SN+T++ ND
        query    NTTWYGNDTWYSNDTWNSNDTWSSND
```

In summary, the HMM-based search is fairly stringent as no significant matches were found to the HIV/SIV envelope model that were not proteins from primate immunodeficiency viruses. Selected examples of envelope matches with weak scores (less than 10) can be found in the accompanying section of Part IV concerned with molecular mimicry. HMM-generated models and database searches were also conducted for Gag, Tat, Vpr-Vpx, Vpu, and Nef, all of which are more evenly fitted than Env to all primate immunodeficiency viruses. Some brief comments regarding the search results follow:

Gag: Significant matches to nonprimate lentiviruses were more common with the HMM consensus for Gag than for Env: Visna, CAEV, EIAV, Jembrana, BIV and FIV all displayed matches, with FIV having the highest score (approximately 85). Most of these matches included the Gag zinc-finger motif, and as a result many matches above a score of 20 were also observed for proteins other than Gag. The tetrahymena cnjB gene product, for instance, scored 35.33, which was comparable to scores for some of the nonprimate lentiviral Gags. Many scores less than 20 were encountered for the zinc-finger motif, for example:

```
         HMM Consensus    RKIIKCFNCGKEGHIARNCRAPRKKGCWKCGKEGH
                          R + KCFNC  EGH    C+ P  +GC  CG  GH
C. elegans RNA helicase   RGPMKCFNCKGEGHRSAECPEP-PRGCFNCGEQGH
```

The log-odds score for this match was 4.67. We conclude from this search that the HMM model was especially good for picking up zinc-finger motifs of a certain kind.The so-called major homology region (MHR), which in the HMM consensus appears as IRQGPKEPFRDYVDRFYKTL, showed up only in matches with lentiviral Gags or with retroviral type D Gag sequences, such as those of SRV.

Tat: Of the nonprimate lentiviruses, only BIV and the Jembrana disease virus possessed significant match scores to the HMM consensus, 43.75 and 37.94 (versus an HIV-1 or HIV-2 score of about 280). These matches were across the second and third domains of the first coding exon, which encompass cysteine residues involved in intramolecular bonding and the R/KKGLGI motif that is thought to constitute the minimal Tat. EIAV displayed a weaker match, 10.43, and only in the second domain. These findings corroborate earlier judgments in the field that only BIV, and possibly EIAV, among nonprimate lentiviruses, possess a "true Tat" (the Jembrana virus has been sequenced subsequent to that conclusion.) Hence, FIV's match in the fourth domain, with a score of merely 5.26, is not considered Tat-specific:

```
HMM Consensus   KKRRQRRRTPQKS
                KK RQRRR ++K+
          FIV   KKKRQRRRRKKKA
```

Vpr-Vpx: Given the paralogous relationship between Vpr and Vpx, one HMM Consensus was generated for the two proteins. The highest score against this consensus for sequences other than primate lentiviral sequences was with SA-OMVV, the Visna relative:

```
HMM Consensus   MEQAPWEfPRERIDQGWEWDPQRE
                ME+A    PR    +G     +RE
    SA-OMVVSA   MEEA----PRR--RPG----GSRE
```

The score was 10.44. A nearly identical score was attained by a ligand for Fas antigen, but clearly due to mere compositional homogeneity:

```
 HMM Consensus   GPGGWRRGPPPRNPPSRSMH
                 GPG+ RR PPP++PP+ S +
ligand for Fas   GPGQ-RR-PPPPPPPP-SPL
```

Vpu: In contrast to other HMM consensus sequences, the discriminator for Vpu was generated solely from HIV-1s. Match scores varied significantly—251 for the BAL strain but merely 45.38 for ANT70—suggesting that the model was overfitted to M group viruses. The subtype D virus ELI had a score of 229. Among non-HIV-1s, the highest score, 10.93, was found for a toxin receptor:

```
             HMM Consensus   MQPLQILAIVALVVAaIIAIVVW
                             +  L+I A V LV ++  A VVW
C5a anaphylatoxin receptor   ILALVIFAVVFLVGVLGNALVVW
```

Rev sequences from SIV displayed weak similarity with the N-terminus of the HMM model for Vpu:

```
HMM Consensus   MQPLQILAI
                +Q LQ LAI
      SIV Rev   IQQLQRLAI
```

Nef: With exception of primate immunodeficiency viral sequences, no database sequence displayed striking similarity to regions of the NEF HMM consensus. When mediocre scores were seen, they invariably resulted from compositional homogenity—cysteines, acidic amino acid residues, etc. This result is in contrast to the widespread claims regarding Nef "homologs" (i.e., similarities) in the literature of HIV.

From these preliminary studies, we conclude that the HMM approach to database searching is sensitive and specific. As a supplement to searches based on an HMM consensus, users should consider searching with COBBLER sequences deduced from the BLOCKMAKER program (http://www.blocks.fhcrc.org). Future improvements to the approach will include the introduction of structural information. We shall now turn to structure-prediction and the interplay between primary sequence alignment and structure-based alignment.

C. PROTEIN STRUCTURE PREDICTION

A promising starting point for predicting a structure for a given amino acid sequence is to determine whether that sequence is sufficiently similar to any other sequence for which biophysical data, ideally X-ray crystallographic data, is available. For sure, sequences that are 50% are more similar will have similar structures, while less similar sequences can have similar folds over core regions. The focus herein will be upon weakly similar sequences, in particular upon those of envelopes for which comparatively little biophysical data, beyond limited NMR, is available.

The earliest structure prediction algorithms, such as the Chou-Fasman algorithm, possess a predictive accuracy of no better than about 55%, partly due to the small set of known structures upon which they depend and partly due to their assumptions. Three-state predictions—helix (*H*), sheet (*E*) and coil (*C*)—are more accurate than four-state predictions that include turns (*T*); the accuracy is poorest at the ends of polypeptides and best in the core regions. Secondary stucture prediction in general is most reliable for transmembrane helices. With the buildup of the protein database and the development of more powerful algorithms, which especially take into account multiple sequence alignments, the predictive accuracy for secondary structure can now reach slightly better than 70%.

SOPM (self-optimised prediction method) is an example of a recent approach to protein secondary structure prediction [16–17]. When applied to 239 dissimilar proteins of known structure, this algorithm yields three-state prediction accuracies of 69% to 73%. On the other hand, because it involves sizeable subdatabases of sequences and their known structures, it will take longer to run than the older, less accurate algorithms. The basic ideas used in the SOPM are as follows.

First, a sliding window of a fixed size is applied to the protein sequence of unknown secondary structure to define a set of overlapping peptides. For example, suppose we are given the sequence KPQRNSKSTAAL . . . with a window whose size is eight amino acids long and which is moved one amino acid over at each step. The resultant set of octapeptides will be KPQRNSKS, PQRNSKST, QRNSKSTA, RNSKSTAA, NSKSTAAL . . . . Note that most of the amino acids of the original sequence will belong to successive octapeptides, each differing from the previous peptide by the removal of an amino acid from one end and the addition of an amino acid to the other.

Next, each of the peptides thus derived from the query sequence is now compared to a database of peptides that has been created by similar means from a database of proteins of known secondary structure. If the peptide from the query sequence matches a peptide from the database above a certain threshold of similarity, then the similarity score is added to the conformational scores for each of the amino acids in the peptide. In our example, suppose that the first peptide KPQRNSKS matches a peptide in the database RPQRDTKS whose known structure is *HHCCCEEE*, and that the similarity score between these two peptides is 30. If this score is above the threshold parameter, then 30 will be added to the first two amino acids' helical conformational scores, to the next three amino acids' coil conformational scores and to the last three amino acids' sheet conformational scores. There may be other peptides in the database matching the query peptide with alternative predictions for the secondary structure of each of the amino acids, and all these predictive scores will be added together in each of the conformational categories, resulting in a distribution of scores over the possible secondary structure conformations. After the first query peptide has been compared, the process will continue for each of the remaining peptides in the query set. The final scores for an amino acid belonging to eight successive query peptides will thus include the scores for the comparisons of all eight of these peptides against the entire database of peptides of known structure.

After all comparisons have been made, each amino acid in the original protein will have values associated with its propensity to adopt a conformation in each of the secondary structure classes. From the method of calculation detailed above it is clear that the empirical evidence for the prediction of the secondary structure of the amino acid weighs most heavily for that class with the highest score.

There are two additional statistics concerning the distribution of the scores over all the classes that can be revealing of the predictive power of this approach. The first is the actual magnitude of the scores for any given amino acid. If these are small relative to the cumulative scores for other amino acids, it may indicate a lack of information for the prediction of the secondary structure conformation of that amino acid. This could happen for two reasons: first, if the amino acid is within the window size to either terminus of the original protein, it will belong to proportionately fewer query peptides and have fewer comparisons with the database that could add to its score; second, the amino acid could belong to a series of peptides that for some reason are poorly represented in the database of known structures, and could thus have few comparisons to the database having a large enough similarity score to be added to the conformational scores for the amino acid. In either case, values that are low in magnitude indicate a lack of information in the database for the amino acid in its given environment.

The second statistic that is pertinent to the predictive value of a set of scores for a given amino acid is the difference between the scores of the highest and next-highest scoring classes of secondary structure. If this difference is small, it may be inferred that the information in the database for the amino acid in this particular environment is conflicting. For example, suppose that approximately half of the peptides contributing to a given amino acid's conformational scores support a helical structure, while the other half support its being classed as an element of a beta sheet. In this scenario, it is likely that the cumulative scores for helix and sheet for this amino acid would be nearly equal, and thus the difference between them would be near zero. In order to make this statistic independent of the magnitudes of the scores (which were accounted for in the former statistic), one may normalize the values by dividing the difference between the highest and next-highest scores by the magnitude of the highest score.

It is the widespread wisdom at this time to evaluate sequences, whenever possible, by more than one algorithmic approach; some methods are better for predicting helices, others for predicting sheets, etc. The SOPMA server (*http://www.ibcp.fr/predict.html*), therefore, submits a sequence to alternative methods of structure prediction—Gibrat, Levin, DPM and the PhD [18–21]—and also generates a consensus over those and the SOPM prediction itself. The HMM-generated "most likely" sequence was submitted to the SOPMA suite, producing predictions using the five individual algorithms as well as a consensus prediction, with the following results(H = helix, E = beta sheet, T = turn, and C = coil; blocks, defined by the MOTIF program in BLOCKMAKER, are indicated).

Although the HMM model has been overfitted in this case to HIV-1, the structures should be somewhat conserved across primate lentiviral boundaries. Gallaher and coworkers have explored this for the surface and transmembrane components of the lentiviral envelope [22,23]. Because the HMM-generated "most likely" sequence embodies information from hundreds of primate immunodeficiency viruses, it offers a reasonable test of their "eclectic" models derived from representative lentiviral sequences [22,23].

The Gallaher model for the surface protein (dependent primarily upon the Chou-Fasman algorithm) identifies five helical regions, all five of which are strongly or moderately predicted by the SOPMA suite: the 1st overlaps the 1st definable block in gp120 (figure 2 and Part II); the 4th is included in the 4th block and the 5th is included in the 5th block. Helices 1, 3 and 5 are strongly evident in HIV-2 sequences such as ROD. Furthermore, SOPMA suggests a small stretch of helix at the C-terminal end of the V3 loop and also at the C-terminal end of the gp120. In general, helices are the most predictable of secondary structures. Turns are weakly predicted following the V2 loop (in the 2nd definable block) and twice within the V3 loop, as we have come to expect. A stronger prediction is for the second turn that follows the third helical region, which separate V3 and V4; this "hinge" coincides with two of the four highly conserved CD4 contact residues and is bordered by the 3rd definable block. The two other contact residues for CD4 interaction occur in the 4th block, which includes the 4th helical region. An examination of predictions for several HIV-1 V3 loop sequences of different subtypes suggests that the Levin, DPM and SOPM methods most consistently predict the putative type II beta turn at the crest of the V3 loop. Further structural analysis of V3 loops is provided by Catasti and Gupta in Part III.

```
                1       10        20        30        40        50        60
                |        |         |         |         |         |         |
most-likely     MRVKGIRRNYQHLWRWGILLLGMLMICSAAENLWVTVYYGVPVWKEATTTLFCASDAKAY
Gibrat method   EEEEEEEEEHHEHEHHHHHHHHHHHHHHHHCCCEEEEEEECCCEEECHCCHHHHHHCCHCCC
Levin method    ESESSCECCSTTTCTCHHHHHHEHEEECCCTTCEEEEEEECCCCCCHCCHEEECCCCCHCH
DPM method      CCEECECCCCCCEEEEEEEEEEEEEEEECHHHHEEEEEEEECECEEHHHCEEEEEHCCHCCC
SOPMA predict   HHTTTCCHHHCTEECCCCEEHCCCHHHHTCCCCHEEEEECCCTTCCCCCCCCCCCHHHHH
PhD method      CCCCCHHHHHHHHHHHHHHHHHHHHHHHEECCCCEEEEEEEEEEECCCCCCEEEEEECCCCCC
Consensus       -CE-C----H---E--HHHHHH-HHHE--CCCEEEEEEEECCC--CHCC-EEE--CCHCCC
                                      ---     alpha-helix 1        -- turn?
most-likely     DTEVHNVWATHACVPTDPNPQEIVLENVTENFNMWKNNMVEQMHEDIISLWDQSLKPCVK
Gibrat method   HHHHHEHEHEECECCCCCCCCHEEEEHHHHHHHHHHHCHHHHHHHHHHHHHHHHCCCCCCC
Levin method    HHHHHHHHHHCCECCCCCCSCSCEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHCHTHCCTTCC
DPM method      CCCEEEEEHECCEECCCCCCCHEEEEEEECCCCTCCCCCEEHHHHHCEEEEECCCCCCCEEC
SOPMA predict   TTTHHHHHHCCEEECCCCCCCEEEEEEEEEEHHHCTTCHHHHHHHHCEEECCCCCHHHHHH
PhD method      CCCCCCCCCCCCCCCCCCCCCCEEEEECCHHHHHHHHHCHHHHHHHHHHHHHHHHHHHCCCCCEE
Consensus       ---HH-H-H-CCECCCCCCCCC-EEEE--HHHHHHHHHCHHHHHHHHHHHHHH-H-CCCC--C
                         -->             V1 loop                 <-- -->
most-likely     LTPLCVTLNCTDVNATNTNNTTNTTKIDMINETSSCIRQDNCTGLEKGEIKNCSFNITTE
Gibrat method   CCCCEECECCCCECCEECCCCCCCEEEEEEECCCCECCCCCCCCCCCCCCEEECCCCCCCHH
Levin method    CCCEEEEEEECECCCHTCTTCCCCCEEEEEEECCCCSECCCSTCCCCCTTCCHHCCHHEEHH
DPM method      ECCEEEEECCCCECCCCCCCCCCCCCCCECEECCCCCEECCCTCCCCTCCCECCTCECEEEE
SOPMA predict   HCCTEEEECCEECCCCCTTCCCCCCEEEECCCCCCCCCCCCCCEEEETTEEEEEECCHEEH
PhD method      CCCCEEEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCECCEEEECCCC
Consensus       CCC-EEEECCCCCCCCCCCCCCCCCCEEEEECCCCCECCCCCCCCCCCCCE---C-C-E-H
                       V2 loop                <--           hinge?
most-likely     IRDKKQKEYALFYKLDVVPIDNNNTSYRLINCNTSVITQACPKVSFEPIPIHYCAPAGFA
Gibrat method   HHHHHHHHHHHHHHHHHEEEECCCCCCEEEEECCCCCCCECCCCCCCCCCCHCECCCCCHH
Levin method    CHCSSHHHHHHHCECCECCCCTCCSEEEEECTCCCEEEEHCCTCCTCSCEEEEECCCTSCE
DPM method      ECCCCTCHHHEEEEECEECECTTTCCEEEECCCCEEEECCCCCECECCCCCCCCCCCCCEH
SOPMA predict   HHHCCHHHHHHHHHHHHHEEECCCCCCCEEECCCCCCCEEEEEEEEEECCCCEEEEECCHHHH
PhD method      CCCHHHHHHHHEECCCEEEECCCCCEEEEEEECCCCEEEECCCCCCCCCCCCEEEECCCCEE
Consensus       -HCCCHHHHHHH---CEEEECCCCCCEEEECCCCCEEEECCCCCC-CCCCEEECCCCC-H
                                              --   alpha helix 2
most-likely     ILKCNDKKFNGTGPCKNVSTVQCTHGIKPVVSTQLLLNGSLAEEEIVIRSENFTDNAKTI
Gibrat method   HHHHCCCCCCCCCCCCCCCCCEEEECCCCCCEEEHHHHHCCCHHHHHEEEHHHCCCHCHHHE
Levin method    EEESCCTSCCTCCCCCCCEEECCCCECCCECETCEEEHHCCCTTCEEEECHCCCHCCSCE
DPM method      EECTTTCCCTTCTCCCTEEEEEEECCECCEEEEEEEECTCCCHHHHEEEECCCCCCCCCCEE
SOPMA predict   HHHHTCCCCCCCCCCCCCCHHECCTTCCCEEEEEEHHHHHHTTHHHEEHHHHHHHHHHHHE
PhD method      EEEECCCCCCCCCCCCCCCEEEEEEECCCCCEEEEEEEEEECCCCCHHHEEEEEEEEECCCCCCEEE
Consensus       EE--CCCCCCCCCCCCCCCEEEEECCCCCCEEEEEEEE--CCCHHHHEEEE-HCCCHCC--E
                                        V3 loop
                           -->  turn?    turn?             <--      -- alpha
most-likely     IVQLNESVEINCTRPNNNTRKSITIGPGQAFYATGDIIGDIRQAHCNISGAKWNETLQQV
Gibrat method   EEEECCCCEEEECCCCCCCCEEEEECCCCEEEECCCHCEEEEEHEECCCCCCHHHHHHHH
Levin method    EEEECSHHCEECCCCTCCSCSCEECCTTCCEEETSECEHECCHCCCCHTTCCCHHHHHHH
DPM method      EEEECCEEEECCCCCTCTCCCEEEECCCCCECCCCCEECCECHHCCCECTCCCCCCEHEE
SOPMA predict   EEEECCCHECCCCCCCCCCCCCEEECCCCTTECCCCCCCCEEHHEEECCCCCHHHHHHHH
PhD method      EEEECCEEEEEECCCCCCCEEEEEEECCCEEEEECCCCCCCHHHCCHHHHCHHHHHHHHHH
Consensus       EEEECC--EEECCCCCCCCCCCEEEECCCC-EEECCCCCCCE-HHC-C-CCCCHHHHHHHH
                                   CD4 CD4
                                     | |
              helix 3 --         hinge?
most-likely     AKKLREQFGNKTIIFNQSSGGDPEITTHSFNCGGEFFYCNTTQLFNSTWNNGTWNSTESN
Gibrat method   HHHHHHHHCCCEEEEECCCCCCCEEEEECECCCCEEEEECCEEEEEEEEEECCCEEEEECCC
Levin method    HHHHHHHTTTSEEEEECCCSCCCECCECSTCCTCCCEEECCTHHECCTCTTTCCCCTCCT
DPM method      HHHHHHCTCCCEEEECTTTTCTCCCCECCCTTCCCEEEECCCCEECCCCTTCCCCCCTTC
SOPMA predict   HHHHHHTTTCCEEEEETTTCCCTEEEEEEEEEEECCEEEEEEEEEEECTTCCTTTCCCCCCC
PhD method      HHHHHHHCCCCEEEEECCCCCCCEEEEEEEEECCCEEEEEEEEHHHHCCCCCCCCCCCCCC
Consensus       HHHHHHHTCCCEEEEECCCCCCCEEEEEECE-CCCCEEEECC--EECC-C-TCCCCCCCCC
                             CD4
                               |                              CD4
                      -- alpha helix 4 --                       |
most-likely     DTITLPCRIKQIINMWQEVGKAMYAPPIEGQITCSSNITGLLLTRDGGDNNSTNETFRPG
Gibrat method   CCEEEHHCCHCEHHCHHHHCCCCCCCCCCCEEEECCCCCEEEEEEECCCCCCCCCEEECCC
Levin method    SCEEEEECHHHHHHHHHHHTSCECCSCCCTCCCCCTTCCEEEEEEETCCCCCTTCCCCCCT
DPM method      CCECCCCCEEEEEEEEECHECCHCCCCCCCCCCEECCCCECCEEECCCTTTTTTCCCCCCTT
SOPMA predict   CCEEEECCHHHEEEHHHHHTEEEECCTHCCCCCCCCCCCHEEHCCCCCCCCCCCCCCCCCC
PhD method      CCCCCCHHHHHHHHHHHHHHHHHCCCCCCCCCEEEEECCCCEEEEEEECCCCCCCCCCCEEECCC
Consensus       CCEEE--CHHHEHHHHHHHH-C-CCCCCCCCCEECCCCCCEEEEEEECCCCCCCCCCCCCCC
```

```
                1        10        20        30        40        50        60
                |        |         |         |         |         |         |
                    --alpha helix 5--              gp120/gp41 fusion domain
most-likely     GGDMRDNWRSELYKYKVVKIEPLGVAPTKAKRRVVQREKRAVGLGAVFLGFLGAAGSTMG
Gibrat method   CCCCCCECHHHHHHHHEHEHCHHHHCCHHHHHHHHHHHHHHHHHHHHHEEEECHCCCEEE
Levin method    CCCCHHHHHHHHHHHCCCECCCCECCCCCHHHHEHHHCCHHHHCHCEEHCCCCCCCCCCC
DPM method      TCTCCCCCCCHECEEEEEEECCCCCCCCCHHHEEEEHHHHHECECEEEEEEECTCCCCCC
SOPMA predict   CCCCCCTHHHHHETTEEEEEECCCCCCHHHHHHHHHHHHHHHHTTCEEEEEEECCCHHHH
PhD method      CCCCHHHHHHHHCCCCEEEEECCCCCCHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHH
Consensus       CCCCCC-HHHHH----EEEECCCCCCCCHHHHHHHHHHHHHHH---EEEEEEC-CCC---
                                  turn? --       extended helix        --
most-likely     AASITLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVLAVERYLKDQQ
Gibrat method   EHEEEHHHHHHHEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
Levin method    CCCCEEECCCCCCCTHCCCCTTHHHHHHHHHHHHHCEEEEHCCHHTHHHHHHHHHCTTCC
DPM method      CCCEEEEEHHHEEEEEECCCCHEHHHHHHHHHEEEEEEEEEECHHHHHEEHEHHEHCCCC
SOPMA predict   HHHHEEEEHHHCCCCHHCCCHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHC
PhD method      HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
Consensus       -H--EEE-HHHH---HH--CHHHHHHHHHHHHHHHHH---HHHHHHHHHHHHHHHHHHHC
                       turn?          turn?         --     extended helix
most-likely     LLGIWGCSGKLICTTTVPWNSSWSNKSLTPIWNNMTWMEWEREIDNYTALIYTLLEESQN
Gibrat method   EEEECEECCCEEEEEEECCCCECCCCCEEEEEHCCCHHHHHHHHHHHHHHEEEHHHHHHH
Levin method    HHHCECCTCCEEETEECCCCCCSTTTTECEHCHHHCHHHHHHCHHHHHHHHHHCCCHCCC
DPM method      EECEETTCCCEEEEEEECCCCCCTTCCCCCCCCCCCEHHHHHHECCCCEEEEEEHHHTCC
SOPMA predict   EEEEEEETTCEEEEEEECCCEECCCCCCCCCCHHHHHHHHHHHHHCCHHHHHHHHCCHHH
PhD method      HHHHCCCCEEEEEEEECCCCCCCCCCCHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHHHH
Consensus       EE-EE--CCCEEEEEEECCCCCCCCCC-C--CHCCCHHHHHHHHHHHHHHHHH-HHHHHH
                    extended helix and transmembrane region           -- turn?
most-likely     QQEKNEQELLELDKWASLWNWFDITNWLWYIKIFIMIVGGLIGLRIVFAVLSIVNRVRQG
Gibrat method   HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEECCCCCEEEEEEEEEEEEEEECC
Levin method    CCCHHHHHHHHHHSCCCCHHHHHHHHHHHHHHHEEEHHTSCCCHHHHEHHHHHHHHCCSS
DPM method      CCTCCCHHHHHHCHCHCEECEECEECEEEEEEEEEEEEECEEEEEEEEEEEEEEEEECCC
SOPMA predict   HHHHHHHHHHHCCCCCCCCETCCCCCEEEEEEEEEEEEECTTCCEEEEEEEEEEEEEEEC
PhD method      HHHHHHHHHHHHCHHHHHHHHHHHHHHHHHHHHHEEEEECEHHHHHHHHHHHHHHHHHHC
Consensus       HHHHHHHHHHHHCHCHC-HHHHHHHHHHHHHHHEEEEEEC--C-EEEEEEEEEEEEE-CC

most-likely     YSPLSFPPGYIQQTHLPAPRGPDRPEGIEEEGGERDRDRSWRLVNGFLALIWDDLRSLCL
Gibrat method   CCCCCCCCCCEEEECCCCCCCCCCCCCCEHCCCCCCHCHHHHHHHHHHHHHHHHHHHHHH
Levin method    CCCCCCCCSSCEECCCCCCCCCCCCHHHHTCCCCHHHHHHHHHHHHHHHHHHSTSCCHEH
DPM method      CCCCCCCCCCECCCCCCCCCTCCCCCCCCTCCCTCCCCCCEEEECCEEEEEECCCCCEEE
SOPMA predict   CCEEECCCCCEEEEEEEEHHTCCCCHHHHHHTTTCCCHTCCCEECCCCEECCCCCEEEEE
PhD method      CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHH
Consensus       CCCCCCCCCCEEECCCCCCCCCCCCCCC-HCCCCCCHCCC---HHHHHHHHHC-C--HEH

most-likely     FSYHRLRDLLLIVARIVELLGRSSLKGLRRGWEALKYLWNLLQYWSQELKNSAVSLLNAT
Gibrat method   HHHHHHHHHHHHHHHHHHHHCHCCHHCEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
Levin method    HETHCHHHHHHHHHHHHHHHCTTCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
DPM method      EEECCEHCEEEEEEEEEEEETTCCTCCTTCCHHHHHEEECEECEECCHTCCCCEEEECHC
SOPMA predict   EETTTTHHHHHHHHHHHHHHTCCCHHHHHHHHHHHHHHHHHHHHHHHHHHTTHEEEECTH
PhD method      HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
Consensus       HE-H-HHHHHHHHHHHHHHH--CCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

most-likely     AIAVAEGTDRVIEVLQRIGRAILHIPRRIRQGLERALL
Gibrat method   HHHHHCCCHHHHHHHHHHHHEEECCCHHHHHHHHHHHH
Levin method    HHHHHTCCCHHHHHHHHHCHHECHCCHSHCTTCHSHHE
DPM method      HEHEHCTCCEEEEEEEEECEEEEEECCEEECCHHHHCC
SOPMA predict   HHHHHCCTTEEEEEEECCCCEEEECCCEHHHHHHTTTT
PhD method      HHHHHHHHHHHHHHHHHHHHHHHHCHHHHHHHHHHHCC
Consensus       HHHHHCCC-HHHHHHHHHCHHEEECCC-HHHHHHHH--
```

The transmembrane portion of envelope contains more helix than does the surface portion, as determined by circular dichroism, and the structure predictions confirm this for the HMM consensus. An extended helix, encompassed by the 7th definable block, immediately preceeds the immunodominant domain and another, encompassed by the 8th block, follows it. Another predictable helix coincides with the membrane spanning domain, as we would expect. A predicted turn associated with the RQGY peptide that ostensibly signals the terminus of the transmembrane region [24] is not supported by the SOPMA suite.

In the future, we hope to refine the HMM analyses of HIVs and SIVs, first, by addressing the overrepresentation problem of HIV-1s and, second, by integrating the structural information with the primary sequence data.

References

[1] Doolittle, R.F. (1987) *Of URFS and ORFS: A Primer on How to Analyze Derived Amino Acid Sequence* University Science Books, Mill Valley, California.

[2] McClure, M.A., Vasi, T.K., and Fitch, W.M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.* **11**: 571–592.

[3] Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M.A. (1994) Hidden Markov models of biological primary sequence information. *Proc. Nat. Acad. Sci. U.S.A.* **91**: 1059–1063.

[4] Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994) Hidden Markov methods in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**: 1501–1531.

[5] Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77:257–286.

[6] McClure, M.A. and Raman, R. (1995). Parameterization studies of hidden Markov models representing highly divergent protein sequences. Proceedings of the 28th Annual Hawaii International Conference on Sysytem Sciences, pp. 184–193, published by the IEEE Computer Society Press

[7] Luthy, R. and Eisenberg, D. (1992). Protein. In *Sequence Analysis Primer* (eds. M. Gribskov and J. Devereux), pp. 78–82. W.H. Freeman and Company, New York.

[8] Gribskov, M. and Veretnik, S. (1996). Identification of sequence patterns with profile analysis. In *Computer Methods for Macromolecular Sequence Analysis* (ed. R.F. Doolittle). pp. 146–159, Academic Press, Inc., San Diego.

[9] Eddy,S. (1995) User's guide for HMMER; Hidden Markov Models of protein and DNA sequence (version 1.8), Washington University of St. Louis, 660 S. Euclid, Box 8232, St. Louis MO 63110

[10] Eddy,S., Mitchison,G., and Durbin,R. (1995) Maximum discrimination hidden Markov models of sequence consensus. J Comp Biol 2: 9-23.

[11] Wills,C., Farmer, A., and Myers, G. (1996) Rapid sequon evolution in human immunodeficiency virus type 1 relative to human immunodeficiency virus type 2. AIDS Res Human Retro 12:1383-1384.

[12] Henikoff,S. and Henikoff,J.G. (1994) Protein family classification based on searching a database of blocks. Genomics 19:97–107.

[13] Henikoff,J.G. and Henikoff,S. (1996) BLOCKS database and its applications. In *Computer Methods for Macromolecular Sequence Analysis*, ed. R.F. Doolittle, pp.88–105, Academic Press, San Diego.

[14] Lamers,S.L., Sleasman,J.W., and Goodenow,M.M. (1996) A model for the alignment of ENV V1 and V2 hypervariable domains from human and simian immunodeficiency viruses. AIDS Res Human Retro 12: 1169-1178.

[15] Myers, G.(1996) Retroviral Sequences. In *Retroviruses*, ed. by Coffin,J. Hughes,S. and Varmus,H. Appendix I. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY.

[16] Geourjon, C. and Deleage, G. (1994) SOPM: a self-optimised prediction method for protein secondary structure prediction. *Prot. Eng.* **7**: 157–164.

[17] Geourjon, C. and Deleage, G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS* **11**: 681–684.

[18] Gibrat, J.-F., Garnier, J., and Robson, B. (1987) Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* **198**: 425–443.

[19] Levin, J.M., Robson, B., and Garnier, J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS* **205**: 303–308.

[20] Rost, B. and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function and Genetics* **19**: 55–72.

[21] Deleage,G. and Roux,B. (1987) An algorithm for protein secondary structure prediction based on class prediction. Prot Eng 1: 289–294.

[22] Gallaher,W.R., Ball,J.M., Garry,R.F., Griffen,M.C., and Montelaro,R.C. (1989) A general model for the transmembrane proteins of HIV and other retroviruses. *AIDS Res Human Retro* **5**:431–440.

[23] Gallaher,W.R., Ball,J.M., Garry,R.F., Martin-Amedee,A.M., and Montelaro,R.C. (1995) A general model for the surface glycoproteins of HIV and other retroviruses. *AIDS Res Human Retro* **11**:191–202.